Stoch Environ Res Risk Assess
DOI 10.1007/s00477-010-0429-5

ORIGINAL PAPER

Clustering and forecasting of dissolved oxygen concentration on a river basin

Marco Costa · A. Manuela Gonçalves

Abstract The aim of this contribution is to combine statistical methodologies to geographically classify homogeneous groups of water quality monitoring sites based on similarities in the temporal dynamics of the dissolved oxygen (DO) concentration, in order to obtain accurate forecasts of this quality variable. Our methodology intends to classify the water quality monitoring sites into spatial homogeneous groups, based on the DO concentration, which has been selected and considered relevant to characterize the water quality. We apply clustering techniques based on Kullback Information, measures that are obtained in the state space modelling process. For each homogeneous group of water quality monitoring sites we model the DO concentration using linear and state space models, which incorporate tendency and seasonality components in different ways. Both approaches are compared by the mean squared error (MSE) of forecasts.

Keywords Hydrological basin · Water quality · Clustering · State space model · Linear model · Kalman filter

M. Costa (✉)
Escola Superior de Tecnologia e Gestão de Águeda, Universidade de Aveiro, Apartado 473, 3754-909 Águeda, Portugal
e-mail: marco@ua.pt

A. Manuela Gonçalves
Departamento de Matemática e Aplicações, Universidade do Minho, Campus de Azurém da Universidade do Minho, 4800-058 Guimarães, Portugal

1 Introduction

As water is a precious asset as well as a potential inducer of riches, water quality monitoring networks are important tools in the management and assessment of surface water quality and they could be improved by means of accurate forecasts of the surface water variables. The European Water Framework Directive (WFD) establishes a common framework for sustainable and integrated management of natural waters. This implies a high level of multi-disciplinary and a tight connection between water management technical bodies and instruments of analysis for decision-making (Vieira 2003).

Surface waters in a river basin are usually submitted to pressures and changes due to human activities. These activities are one of the most important causes of the degradation of water quality, which could pose risks for public health. At a river basin scale there is a need to establish a methodology for systematic data monitoring for the characterization of surface water quality and for the correct analysis of collected data, so that the present and expected future pressures can be identified and understood. Assessment of pressure-state impact interaction can be facilitated using environmental indicator tools (Oliveira et al. 2005).

In this work, we intend to contribute to the discussion and understanding of an environmental issue of such importance to the community, as is the case of the quality control of the surface water of River Ave basin. The River Ave hydrological basin is located in the northwest of Portugal, with an approximate basin area of 1390 km^2 and its main stream length of 101 km (Fig. 1).

In a region such as the Ave valley, with its economic ground highly dependent on industry (predominantly textile: there are about 340 registered factories), water plays

Journal : Large 477 | Dispatch : 20-7-2010 | Pages : 13
Article No. : 429 | ☐ LE | ☐ TYPESET
MS Code : SERRA_840 | ☑ CP | ☑ DISK

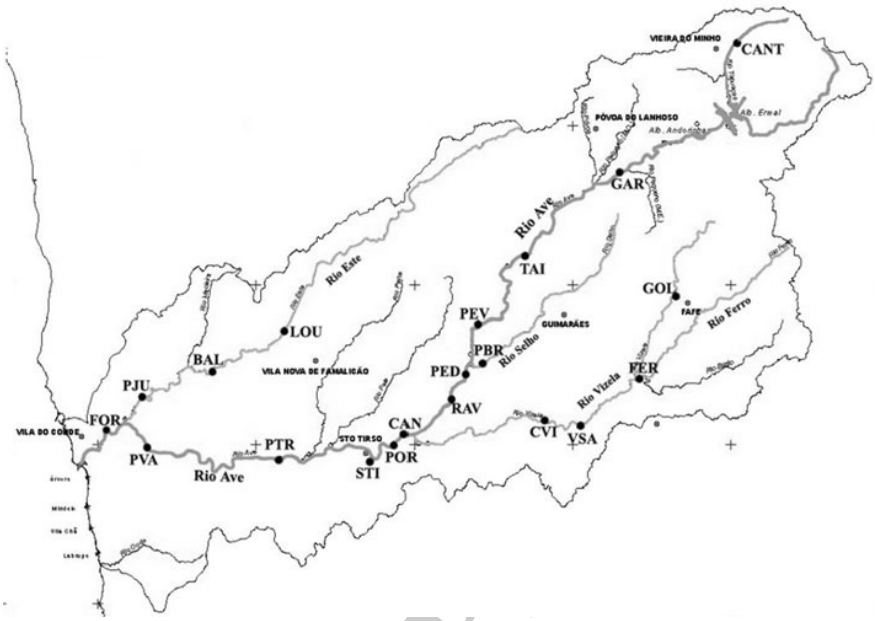

Fig. 1 River Ave hydrological basin located in the northwest of Portugal and its water monitoring sites

undoubtedly a determining role in the industry dissemination in this valley. This region's water streams have been in a situation of obvious environmental degradation for many years. From the 1970s onwards, the worsening of the environmental situation of this basin has led the government authorities to be concerned with the increase of the water pollution in this basin. Since 1988, and as part of a national plan, the Central Administration, through the Regional Directory for the Northern Environment and Natural Resources (DRAN), and the Institute of Water (INAG) monitored the quality of surface water periodically (monthly) along the River Ave and its main adjacent streams. The project PGIRH/N (1988) established a first integrated management plan to the River Ave basin including public organizations such as the Ministry of Public Works, Transport and Communications, LNEC (National Laboratory for Civil Engineering), the Ministry of Planning and Territorial Administration and Coordination Committee of the Northern Region (CCRN). Hydrologic problems had been firstly discussed in the Project of Integrated Management of Northern Water Resources (PGIRH/N and NATO-RIVERS 1994) in co-partnership with the Environmental and Natural Resources Office, DRAN and INAG. Besides, these projects characterize the hydrological Ave basin concerning climatic and orographic properties.

The aim of this study is to identify homogeneous regions based on similarities in the temporal dynamics of variables of water quality measured patterns by observing hydrological series, namely the dissolved oxygen (DO) concentration variable, in mgO_2/l, recorded in time and space, in the River Ave basin with the purpose to evaluate the surface water quality and to obtain accurate predictions.

Clustering has been widely used in environmental problems, namely in climatic themes (Zhu and El-Shaarawi 2009; Gong and Richman 1995; Fovell and Fovell 1993), and atmospheric science (Bengtsson and Cavanaugh 2008; Stone 1989). The combination of state space models with clustering procedures improved the fitting and prediction accuracy. Indeed, the state space approach with Kalman filter technique produces good results in weather radar calibration (Alpuim and Barbosa 1999; Brown et al. 2001), in correct temperature forecasts (Libonati et al. 2008) and in forecasting near-surface parameters (Galanis and Anadranistakis 2002; Boi 2004). The application of different multivariate statistical techniques, such as cluster analysis, helps in the interpretation of complex data matrices to better understand the water quality and ecological status of the studied systems; it also allows the identification of possible factors/sources as well as rapid solutions to pollution problems, and it is useful in verifying temporal and spatial variations caused by natural and anthropogenic factors linked to seasonality. Cluster analysis allows the grouping of river water samples based on similarities in physical–chemical composition (Shrestha and Kazama 2007). Hierarchical agglomerative clustering by the Ward's method was selected for sample classification.

We present a comparative study based on DO concentration considering two approaches: state-space and linear models, both associated to clustering techniques. We identify homogeneous regions, based on similarities in the temporal dynamics of water quality variables measured patterns, following a similar strategy adopted in Bengtsson and Cavanaugh (2008). For each cluster, we establish

Journal : Large 477 | Dispatch : 20-7-2010 | Pages : 13
Article No. : 429 | LE | TYPESET
MS Code : SERRA_840 | CP | DISK

linear and state space models aiming at modelling and forecasting water quality variables. The state-space models, associated with the Kalman filter, allow us to model the studied variable by establishing a dynamic model, where the dependence structure is modelled by a latent state variable. The linear models contain a term for the global trend and the seasonal variation throughout the year. We discuss the quality of the predictions produced by the two approaches by comparing them, via the mean squared error (MSE), in a period of time.

2 Data set description

The Northern Regional Directory for the Environment and Natural Resources and the Portuguese Institute of Water have been collecting various water quality variables (monthly physical–chemical and microbiological analysis) from 20 quality monitoring sites in the River Ave basin.

Although there are more water quality variables available, we selected the DO concentration due to its continuity in measurement at all selected water quality monitoring sites and its importance in the evaluation of the water quality of this river (point sources: industry, domestic wastewater, agriculture, wastewater treatment plants). The DO concentration analysis measures the amount of gaseous oxygen (O_2) dissolved in an aqueous solution. Oxygen gets into water by diffusion from the surrounding air, by aeration (rapid movement), and as a waste product of photosynthesis. The DO in water is one of the most important quality variables to assess the degree of pollution existent in the surface waters of a river's hydrological basin. Low values indicate bad water quality. Organic pollution is the most common type of pollution in this basin and, consequently, a frequent problem is a deficit of DO. This results in anaerobiosis situations, which produce bad smells and destroys organic life. This problem is aggravated by the existence of a sequence of small dams in the River Ave and its main adjacent rivers, which limit the oxygen transfer by aeration.

In this study we consider data series from 16 water monitoring sites because the remaining four monitoring sites are very recent, and data series are very short. Thus, the data sets of 16 water quality monitoring sites of DO concentration used in this work have been monthly measured between 1988 and 2006, with some missing data. Summary statistics for DO in all monitoring sites are given in Table 1.

Data series were separated in two independent data sets: one for the modelling process and the other for the forecast procedure. The choice of these two data sets is established to guarantee a significant data set for the parameters estimation, whereas the remaining data must be sufficient to assess the forecasts accuracy. Indeed, a substantial data set is necessary to fit state space models to water monitoring sites series, while a data set is left with an expressive number of observations in order to properly evaluate the quality of forecasts. Taking this into consideration, for the modelling process we consider data until December 1998 in *Garfe* (GAR), *Ponte Junqueira* (PJU), and *Caldas de Vizela* (CVI) until September 1999 in *Portos* (POR), until January 2000 in *Ponte Brandão* (PBR), *Caniços* (CAN), *Formariz* (FOR) and *Ponte Velha do Ave* (PVA). In the remaining water monitoring sites we consider data until September 2004. Data not considered in the modelling process, namely in the clustering procedure, was relegated for the forecast and assessment stage.

3 Clustering analysis

To identify homogeneous groups of water monitoring sites based on similarities in the temporal dynamics we select the modelling data sets through state space models and Kullback information measure, adapting the methodology adopted in Bengtsson and Cavanaugh (2008). As the DO concentration shows much diversity on tendency and seasonality components in the River Ave and in its main adjacent rivers, we want to identify homogenous clusters of water monitoring sites in the sense of the magnitude of DO concentration, and adopt a simple univariate state space model for each location which considers the DO in its true magnitude. Using a discrepancy measure suggested in Bengtsson and Cavanaugh (2008), we obtain a discrepancy matrix that allows us to identify homogenous groups by applying clustering techniques.

3.1 State space model

State space models have been used in many different areas to describe the evolution of dynamic systems. Such models are defined by the equations

$$\mathbf{Y}_t = \mathbf{H}_t\mathbf{X}_t + \mathbf{e}_t, \quad (1)$$

$$\mathbf{X}_t = \mathbf{\Phi}\boldsymbol{X}_{t-1} + \boldsymbol{\varepsilon}_t. \quad (2)$$

Equation 1 is called the measurement equation and relates the $n \times 1$ vector of observable variables, $\mathbf{Y}_t$, with the $m \times 1$ vector of unobservable variables, $\mathbf{X}_t$, called states. The $n \times m$ matrix $\mathbf{H}_t$ is a matrix of known coefficients and $\mathbf{e}_t$ is a white noise $n \times 1$ vector, called the measurement error, with covariance matrix $E(\mathbf{e}_t\mathbf{e}_t') = \Sigma_{\mathbf{e}}$. Furthermore, the vector of states $\mathbf{X}_t$ varies in time according to Eq. 2, the transition or state equation. In this, $\mathbf{\Phi}$ is an $m \times m$ matrix of autoregressive coefficients and $\boldsymbol{\varepsilon}_t$ is a white noise $m \times 1$ vector with covariance matrix

Journal : Large 477 | Dispatch : 20-7-2010 | Pages : 13
Article No. : 429 | LE | TYPESET
MS Code : SERRA_840 | CP | DISK

Table 1 Sample characteristics for the 16 water monitoring sites

Monitoring site	CANT	GAR	TAI	PBR	RAV	CAN	POR	STI
No. of records	111	124	122	135	99	104	59	101
Start	Sep-93	Oct-89	Nov-92	Oct-88	May-98	Oct-88	Jan-93	May-98
End	Oct-06	Feb-00	Oct-06	Jan-00	Oct-06	Jan-00	Sep-99	Oct-06
Annual mean	9.997	9.403	9.503	7.327	8.230	8.282	7.917	7.621
Monitoring site	PTR	PVA	FOR	PJU	GOL	FER	VSA	CVI
No. of records	100	116	65	95	124	125	124	90
Start	May-98	Oct-88	Jan-93	Oct-90	Nov-92	Nov-92	Nov-92	Oct-90
End	Oct-06	Jan-00	Jan-00	Jan-00	Out-06	Sep-06	Oct-06	Jan-00
Annual mean	7.480	8.459	8.069	9.016	9.687	9.752	9.837	9.211

Monitoring sites: *Cantelães*, CANT; *Garfe*, GAR; *Taipas*, TAI; *Riba d' Ave*, RAV; *Caniços*, CAN; *Portos*, POR; *Santo Tirso*, STI; *Ponte Trofa*, PTR; *Ponte Velha do Ave*, PVA; *Formariz*, FOR; *Ponte Brandão*, PBR; *Ferro*, FER; *Golães*, GOL; *Vizela Santo Adrião*, VSA; *Caldas de Vizela*, CVI; *Ponte Junqueira*, PJU

$E(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t') = \Sigma_{\mathbf{e}}$. The disturbances $\mathbf{e}_t$ and $\boldsymbol{\varepsilon}_t$ are assumed to be uncorrelated, that is, $E(\mathbf{e}_t\boldsymbol{\varepsilon}_s') = \mathbf{0}$, for all t and s. One class of models with particular interest arises when the state vector is a stationary process with mean $E(\mathbf{X}_t) = \boldsymbol{\mu}$.

We fit the monthly DO data using an univariate state space model with constant coefficients ($H_t = H$, in this case we took $H_t = 1$). For water monitoring site i the model represents the observed monthly measure of the DO as a sum of the DO's true value and a white noise term. The true value is denoted by $X_{i,t}$ and is the latent process, and the white noise component by e_{it}. Thus, with $Y_{i,t}$ representing the observed DO concentration for water monitoring site i and month t, we have the observation equation $Y_{i,t} = X_{i,t} + e_{i,t}$, where $e_{i,t}$ is an i.i.d. gaussian zero-mean white noise process with variance $\sigma^2_{e_i}$, i.e., $e_{i,t} \sim N\left(0, \sigma^2_{e_i}\right)$. In their work Bengtsson and Cavanaugh (2008) considered an additive structural state space model with a monthly mean, a seasonal component, a monthly anomaly and a noise term to model the monthly temperature from locations across Colorado, USA. They include these components because their main interest is to perform a clustering process relatively to each structural component. However, this approach can originate different clusters according to the structural component in question. In our case, we want to identify clusters of data series according to a global stochastic behaviour, as in the case of the pseudo-distance used in next sections. Consequently, we intend to focus on the true greatness of DO concentration, because this value indicates water quality. So, we establish a simple model to catch DO concentration magnitude. Thus, we can consider that we are observing the true value of the DO concentration added to a random error due to measurement devices and uncontrolled physical conditions.

For simplicity, we consider that states $X_{i,t}$ are modelled by stationary AR(1) processes, $X_{i,t} - \mu_i = \phi_i(X_{i,t-1} - \mu_i) + \varepsilon_{i,t}$, where $\varepsilon_{i,t}$ is an i.i.d. gaussian zero-mean white noise process with variance $\sigma^2_{\varepsilon_i}$, i.e., $\varepsilon_{i,t} \sim N\left(0, \sigma^2_{\varepsilon_i}\right)$. Some authors (for instance Alpuim and Barbosa 1999) indicate that some types of environmental variables could deviate from the normal curve and, in this case, the assumption of normality to the errors would not be a good choice. However, as will be shown later, the normal distribution seems to fit data and this fact allows implementing gaussian maximum likelihood estimation procedures. Thus, an univariate state space process $Y_{i,t}$ of the DO concentration, at location i and month t, is represented into the state space representation:

$$Y_{i,t} = X_{i,t} + e_{i,t}, \tag{3}$$

$$X_{i,t} - \mu_i = \phi_i(X_{i,t-1} - \mu_i) + \varepsilon_{i,t}. \tag{4}$$

We assume that error measurements $e_{i,t}$ are uncorrelated to the state errors, i.e., $\mathrm{cov}(e_{i,t}, \varepsilon_{j,s}) = 0$, for all i, j, t and s. It is common to consider that the model assumes a prior distribution for $X_{i,0}$ with $E(X_{i,0}) = \mu_{i,0}$ and $V(X_{i,0}) = \sigma^2_{i,0}$, assuming that $X_{i,0}$ is uncorrelated to $\varepsilon_{i,t}$ and $e_{i,t}$ for all t. However, as the process $X_{i,t}$ is a stationary AR(1), we can reduce the number of unknown parameters to be estimated, considering that $E(X_{i,0}) \equiv E(X_i) = \mu_i$ and $V(X_{i,0}) \equiv V(X_i) = \sigma^2_i/\left(1 - \phi^2_i\right)^{-1}$. For each water monitoring site i, we will let $\boldsymbol{\Theta}_i = \left\{\mu_i, \phi_i, \sigma^2_\varepsilon, \sigma^2_e\right\}$ denote the set of parameters for the model (3)–(4) to be estimated by gaussian maximum likelihood.

3.2 Parameter estimates

As we referred in the last section, we obtain the parameter estimates by maximum likelihood estimation assuming gaussian errors. In order to provide the usual iterative procedure to obtain these estimates, we need to introduce

Journal : Large 477 | Dispatch : 20-7-2010 | Pages : 13
Article No. : 429 | LE | TYPESET
MS Code : SERRA_840 | CP | DISK

the Kalman filter algorithm. The main goal of the Kalman filter algorithm is to find estimates of unobservable variables, based on related observable variables through a state space representation in form (1)–(2). Briefly, the Kalman filter is an iterative algorithm that produces an estimator of the state vector $\mathbf{X}_t$ at each time t, which is given by the orthogonal projection of the state vector onto the observed variables up to that time.

Thus, let $\hat{\mathbf{X}}_{t|t-1}$ represent the estimator of $\mathbf{X}_t$ based on the information up to time $t-1$, that is, based on $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_{t-1}$, and let $\mathbf{P}_{t|t-1}$ be its MSE matrix. As the orthogonal projection is a linear estimator, the predictor for the next variable, $\mathbf{Y}_t$, is given by

$$\hat{\mathbf{Y}}_{t|t-1} = \mathbf{H}_t \hat{\mathbf{X}}_{t|t-1}.$$

When, for time t, $\mathbf{Y}_t$ is available, the prediction error or innovation, $\boldsymbol{\eta}_t = \mathbf{Y}_t - \hat{\mathbf{Y}}_{t|t-1}$, is used to update the estimate of $\mathbf{X}_t$ through the equation

$$\hat{\mathbf{X}}_{t|t} = \hat{\mathbf{X}}_{t|t-1} + \mathbf{K}_t \eta_t,$$

where $\mathbf{K}_t$ is called the Kalman gain matrix and is given by $\mathbf{K}_t = \mathbf{P}_{t|t-1}\mathbf{H}_t'\left(\mathbf{H}_t\mathbf{P}_{t|t-1}\mathbf{H}_t' + \Sigma_{\mathbf{e}}\right)^{-1}$. Furthermore, the MSE of the updated estimator $\hat{\mathbf{X}}_{t|t}$ verifies the relationship $\mathbf{P}_{t|t} = \mathbf{P}_{t|t-1} - \mathbf{K}_t\mathbf{H}_t\mathbf{P}_{t|t-1}$. In turn, for time t, the forecast for the state vector $\mathbf{X}_{t+1}$ is given by the equation $\hat{\mathbf{X}}_{t+1|t} = \boldsymbol{\Phi}\hat{\mathbf{X}}_{t|t}$ and its MSE matrix is $\mathbf{P}_{t+1|t} = \boldsymbol{\Phi}\boldsymbol{P}_{t|t}\boldsymbol{\Phi}' + \Sigma_{\mathbf{e}}$.

This recursive process needs initial values for the state vector, $\mathbf{X}_{1|0}$, and for its MSE, $\mathbf{P}_{1|0}$ that will be seen later in more detail. As usual, the orthogonal projection corresponds to the best linear unbiased predictor. When the disturbances $\mathbf{e}_t$ and $\boldsymbol{\varepsilon}_t$ are normally distributed, the state vector and the observed variables are also normal. Therefore, in this case, the orthogonal projection is also the conditional mean value and the Kalman filter is optimal.

For a state space model (1)–(2), under the assumption of normality, the log-likelihood of a sample $(\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n)$ can be written through conditional distributions, that is

$$\log L(\boldsymbol{\Theta}; \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{n}\log(|\boldsymbol{\Omega}|_t) - \frac{1}{2}\sum_{t=1}^{n}\boldsymbol{\eta}_t'\boldsymbol{\Omega}_t^{-1}\boldsymbol{\eta}_t$$

where $\boldsymbol{\Omega}_t = \mathbf{H}_t\mathbf{P}_{t|t-1}\mathbf{H}_t' + \Sigma_{\mathbf{e}}$. Thus, it is possible to obtain the maximum likelihood estimates by maximizing the log-likelihood in order to obtain the unknown parameters by using numerical algorithms, namely the EM algorithm (Dempster et al. 1977) or the Newton–Raphson algorithm (Harvey 1996). In view of the construction of the discrepancy matrix to discrimination and clustering of water monitoring sites, and according to Shumway and Stoffer (1982), we implemented the EM algorithm procedure to the parameters estimation. Taking initial values of parameters $\boldsymbol{\Theta}^{(0)}$—for instance obtained by distribution-free estimators (Costa and Alpuim 2010)—we implemented the iterative procedure of the EM algorithm (Shumway and Stoffer 2006) composed by the following equations

$$\hat{\boldsymbol{\mu}}^{(k)} = n^{-1}\left(\mathbf{I} - \hat{\boldsymbol{\Phi}}^{(k-1)}\right)^{-1}\left(\sum_{t=1}^{n}\hat{\mathbf{X}}_{t|n} - \hat{\boldsymbol{\Phi}}^{(k-1)}\sum_{t=1}^{n}\hat{\mathbf{X}}_{t-1|n}\right);$$

$$\hat{\boldsymbol{\Phi}}^{(k)} = \mathbf{S}_{10}\mathbf{S}_{00}^{-1}; \quad \hat{\Sigma}_{\boldsymbol{\varepsilon}}^{(k)} = n^{-1}\left(\mathbf{S}_{11} - \mathbf{S}_{10}\mathbf{S}_{00}^{-1}\mathbf{S}_{10}'\right) \quad \text{and}$$

$$\hat{\Sigma}_{\mathbf{e}}^{(k)} = n^{-1}\sum_{t=1}^{n}\left[\left(\mathbf{Y}_t - \mathbf{H}_t\hat{\mathbf{X}}_{t|n}\right)\left(\mathbf{Y}_t - \mathbf{H}_t\hat{\mathbf{X}}_{t|n}\right)' + \mathbf{H}_t\mathbf{P}_{t|n}\mathbf{H}_t'\right]$$

where $\mathbf{S}_{00} = \sum_{t=1}^{n}\left[\left(\hat{\mathbf{X}}_{t-1|n} - \boldsymbol{\mu}\right)\left(\hat{\mathbf{X}}_{t-1|n} - \boldsymbol{\mu}\right)' + \mathbf{P}_{t-1|n}\right]$, $\mathbf{S}_{10} = \sum_{t=1}^{n}\left[\left(\hat{\mathbf{X}}_{t|n} - \boldsymbol{\mu}\right)\left(\hat{\mathbf{X}}_{t-1|n} - \boldsymbol{\mu}\right)' + \mathbf{P}_{t,t-1|n}\right]$ and $\mathbf{S}_{11} = \sum_{t=1}^{n}\left[\left(\hat{\mathbf{X}}_{t|n} - \boldsymbol{\mu}\right)\left(\hat{\mathbf{X}}_{t|n} - \boldsymbol{\mu}\right)' + \mathbf{P}_{t|n}\right]$ are computed considering the estimate $\boldsymbol{\Theta}^{(k-1)}$ and Kalman smoothers $\hat{\mathbf{X}}_{t|n}$ and $\hat{\mathbf{X}}_{t-1|n}$ and its MSE, Shumway and Stoffer (1982).

Table 2 shows parameters estimates (for the 16 water monitoring sites in the study) obtained by gaussian maximum likelihood estimation by using EM algorithm, as shown before. In the parameter estimation process, we opted for replacing missing values with seasonal coefficient estimates, because a significant number of missing values can difficult the convergence of parameter estimation process. On the one hand, the global means estimates indicate that monitoring sites *Cantelães* (CANT), *Visela Santo Adrião* (VSA), *Ferro* (FER) and *Golães* (GOL) present the best water quality, considering the DO concentration. On the other hand, at monitoring sites *Ponte Brandão* (PBR), *Ponte Trofa* (PTR), *Santo Tirso* (STI) and *Portos* (POR) we obtained the lowest values means of DO concentration, i.e., these locations have the worst water quality. These conclusions are reinforced by the analysis of error variances estimates. Indeed, low values of both variances estimates, mainly the variance error estimates of the state equation, are obtained at locations with high means of DO concentration. If we consider the geographical locations corresponding to the lowest values of variance errors estimates, we conclude that DO has less variability at water monitoring sites closer to the sources of River Ave and its adjacent streams.

Residuals analysis was performed for each of the 16 monitoring sites in order to evaluate the models adjustment. Globally, satisfactory fits were obtained in all monitoring sites. Although some histograms indicate slightly skewed residuals, the Smirnov–Kolmogorov test does not reject the normal distribution of residuals in none of the monitoring sites under study. In some monitoring sites, for instance *Ponte Trofa* (PTR) and *Portos* (POR), PACF and ACF plots seem to indicate the existence of a seasonal component. However, as our main objective in this context

Journal : Large 477 | Dispatch : 20-7-2010 | Pages : 13
Article No. : 429 | LE | TYPESET
MS Code : SERRA_840 | CP | DISK

Table 2 Gaussian maximum likelihood parameter estimates for the 16 water monitoring sites in the clustering process

Site	CANT	GAR	TAI	PBR	RAV	CAN	POR	STI	PTR	PVA	FOR	PJU	GOL	FER	VSA	CVI
$\hat{\mu}$	10.017	9.256	9.427	7.240	8.166	8.286	7.843	7.609	7.439	8.457	8.034	9.016	9.642	9.672	9.809	9.357
$\hat{\phi}$	0.568	0.581	0.657	0.537	0.616	0.586	0.646	0.561	0.521	0.644	0.598	0.567	0.596	0.566	0.605	0.645
$\hat{\sigma}_e^2$	0.035	0.271	0.0005	2.199	0.002	0.616	2.203	0.003	0.020	0.194	0.010	0.033	0.008	0.004	0.003	0.002
$\hat{\sigma}_\varepsilon^2$	0.724	1.381	0.780	4.184	2.055	3.136	3.340	4.210	3.517	2.489	2.718	1.240	0.752	0.832	0.825	1.366

is to model the global behaviour to identify homogenous groups concerning DO magnitude, we thought that this aspect could be neglected at this moment, considering the gain in simplicity and parsimony of models.

3.3 Discrepancy measure and clustering results

Adapting the discrepancy measure suggested in Bengtsson and Cavanaugh (2008), we define a discrepancy measure based on the state component by considering the Kullback information (Kullback 1968) for the state densities $f(X|\boldsymbol{\Theta}_i)$ and $f(X|\boldsymbol{\Theta}_j)$

$$d^X(Y_i,\boldsymbol{\Theta}_i;\boldsymbol{\Theta}_j) = E_i \ln\frac{f(X_i|\boldsymbol{\Theta}_i)}{f(X_i|\boldsymbol{\Theta}_j)} = \int \ln\frac{f(X_i|\boldsymbol{\Theta}_i)}{f(X_i|\boldsymbol{\Theta}_j)} f(X_i|Y_i,\boldsymbol{\Theta}_i)\,dX. \quad (5)$$

The pseudo-distance between two monitoring sites i and j, based on (5), defined as a form of the J-divergence (Kullback 1968), accounts for the different lengths of each series of data sets by averaging over time

$$\bar{J}^X(Y_i,\boldsymbol{\Theta}_i;Y_j,\boldsymbol{\Theta}_j) = N_i^{-1}d^X(Y_i,\boldsymbol{\Theta}_i;\boldsymbol{\Theta}_j) + N_j^{-1}d^X(Y_j,\boldsymbol{\Theta}_j;\boldsymbol{\Theta}_i).$$

Employing output from the EM algorithm, including the maximum likelihood estimates, the sample $\bar{J}^X$-divergence (Bengtsson and Cavanaugh 2008) reduces to

$$\bar{J}^X\left(Y_i,\hat{\boldsymbol{\Theta}}_i;Y_j,\hat{\boldsymbol{\Theta}}_j\right) = \frac{1}{2N_j\hat{\sigma}_{\varepsilon_i}^2}\left(S_{11}^{(j)} - 2\hat{\phi}_i S_{10}^{(j)} + \hat{\phi}_i^2 S_{00}^{(j)}\right) + \frac{1}{2N_i\hat{\sigma}_{\varepsilon_j}^2}\left(S_{11}^{(i)} - 2_j\hat{\phi}S_{10}^{(i)} + \hat{\phi}_j^2 S_{00}^{(i)}\right) - 1$$

where smoothing quantities $S_{11}^{(k)}$, $S_{10}^{(k)}$, $S_{00}^{(k)}$ and parameters estimates $\hat{\boldsymbol{\Theta}}_k$ are computed based on the model $(Y_k,\boldsymbol{\Theta}_k)$.

The defined symmetric J-divergence is based solely on the state process and targets only the state densities $f(X_i|\boldsymbol{\Theta}_i)$ and $f(X_j|\boldsymbol{\Theta}_j)$, where i and j are two water monitoring sites and $d^X(Y_i,\boldsymbol{\Theta}_i;\boldsymbol{\Theta}_j)$ provides an unbiased estimate of the Kullback information between $f(X_i|\boldsymbol{\Theta}_i)$ and $f(X_j|\boldsymbol{\Theta}_j)$. The evaluation of the discrepancy measure solely depends on the parameters estimates and on partial values obtained from the parameters estimation procedure associated to each location's model. Thus, as occurs in this case study, the data series does not necessarily have to be reported over the same timeframe. However, the application of any process of data series clustering relative to contemporary periods should be accompanied by an analysis of the impact on results. The legitimacy of the application of this procedure to these data series follows the fact that data do not present an accentuated tendency nor structural changes over time. Moreover, their magnitudes are too small, even in the cases where linear models indicate a statistical significant slope in the linear tendency (as it is shown later); consequently, state space models can accommodate this component through their known versatility and dynamics.

By using the parameters estimates of Table 1 and the partial results of EM algorithm, the calculation of sample values $\bar{J}^X\left(Y_i,\hat{\boldsymbol{\Theta}}_i;Y_j,\hat{\boldsymbol{\Theta}}_j\right)$, $i,j = 1,\ldots 16$ allowed to obtain a matrix of pseudo-distances. In Fig. 2, nearest neighbours are represented based on pseudo-distance matrices. It can be seen from the plot that the location of each water monitoring site in the river or its adjacent streams (closer to the source of the river or to the confluence of the River Ave with its adjacent streams) is an important factor to determine the nearest neighbour. However, the location of point sources such as industries and domestic wastewater induces some neighbour relations between *Santo Tirso* (STI) and *Ponte Brandão* (PBR).

In order to identify potential clusters, we apply clustering procedures by using the discrepancy matrix produced by evaluation of $\bar{J}^X\left(Y_i,\hat{\boldsymbol{\Theta}}_i;Y_j,\hat{\boldsymbol{\Theta}}_j\right)$, $i,j = 1,\ldots 16$. The discrepancy matrix was subjected to Ward's, single linkage and complete linkage clustering procedures (Everitt et al. 2001). Because the three methods produce similar results, we only discussed the results of Ward's method. As seen in the dendrogram in Fig. 3, the identified clusters are given by sites: Cluster I {CANT, TAI, GOL, FER, VSA}; Cluster II {GAR, PJU, CVI}, Cluster III {RAV, PVA, FOR} and Cluster IV {PBR, CAN, POR, STI, PTR}, which are geographically represented in Fig. 4.

Considering the estimates of the processes mean obtained in the estimation procedure and indicated in Table 2, it is clear that the clustering procedure performs a classification of the monitoring sites into a possible water quality scale, in what concerns the annual mean DO concentration. In fact, the estimates of the processes mean in

Journal : Large 477 Dispatch : 20-7-2010 Pages : 13
Article No. : 429 LE TYPESET
MS Code : SERRA_840 CP DISK

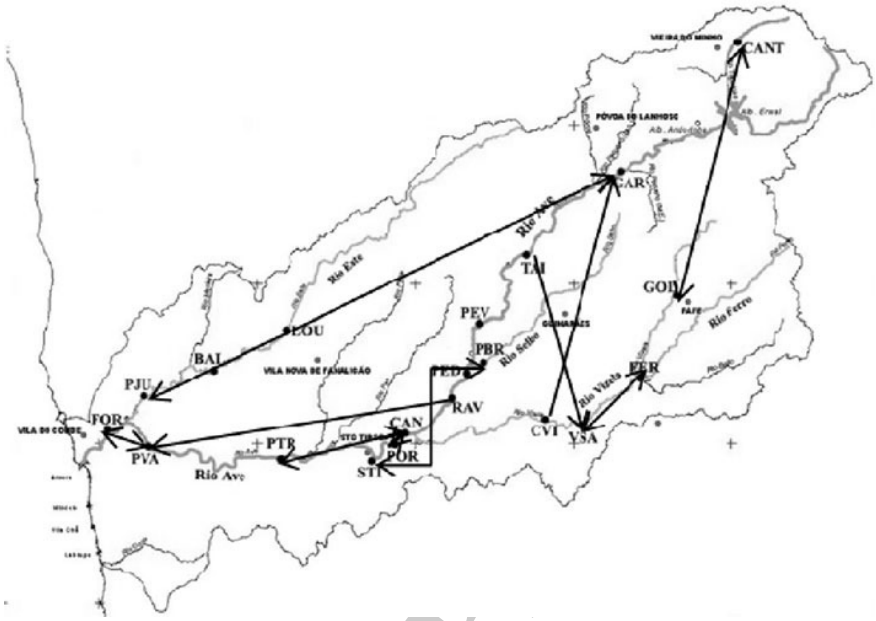

Fig. 2 Nearest neighbours based on pseudo-distance matrix where neighbours are indicated by arrows (TAI → VSA: the nearest neighbour of TAI is VSA)

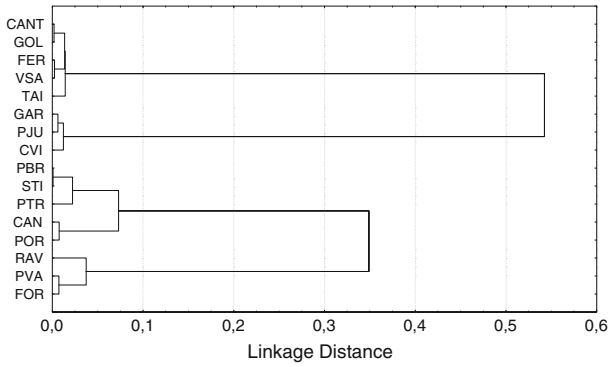

Fig. 3 Dendrogram showing clustering monitoring sites according to DO characteristics based on Ward's method

Cluster I monitoring sites present the highest values obtained from the DO concentration. Indeed, the five monitoring sites of Cluster I present the best water quality annual indicators, while the worst indicators are observed in Cluster IV monitoring sites. On the one hand this methodology allows classifying the water monitoring sites, in regard to annual mean DO concentration, in four categories: from best water quality (Cluster I) to worst water quality (Cluster IV). On the other hand, clustering procedure performs a discrimination of water monitoring sites based on state noise variance. Indeed, Cluster I corresponds to locations with the lowest state noise variances; Cluster II has state noise variances greater than Cluster I and so on. As expected, similar results are not obtained by observing equation variance because the J-divergence $\bar{J}^X\left(Y_i, \hat{\boldsymbol{\Theta}}_i; Y_j, \hat{\boldsymbol{\Theta}}_j\right)$ is based on the state process. Since the discrepancy measure tends to compare state densities $f(X_i|\boldsymbol{\Theta}_i)$ and $f(X_j|\boldsymbol{\Theta}_j)$, it is natural that clustering procedure depicts some patterns on the parameters of these distributions. However, in other studies the order relation in mean and in state noise variance cannot be the same and therefore the results are not similar. Thus, the state space approach, associated to this discrepancy measure, has the merit of allowing the comparison of statistical distribution and does not take into account only one location or scale parameter.

It is interesting to note that Cluster IV has water monitoring sites located downstream the confluences of rivers Selho and Vizela, i.e., where the River Ave receives highly polluted waters of these adjacent streams. The water monitoring sites located in this middle stretch of the River Ave are much more polluted, probably because they are close to densely populated areas with high industrial production units.

4 Forecasting models

As mentioned above, data not considered in the modelling process was used in the assessment of the performance of the adjustment models: LM and state space models, considering the obtained clustering results concerning the mean square error forecast. Thus, for each of the four clusters identified in last section, two models are established by using the available data until September 2006.

As we considered that inside each cluster there is a homogenous annual mean behaviour of DO, it is reasonable to assume that monthly measurements of water monitoring sites inside a cluster are observations of the same process. Therefore, models have to accommodate replicates at each time. Moreover, as the goal at this stage is to

Journal : Large 477 | Dispatch : 20-7-2010 | Pages : 13
Article No. : 429 | LE | TYPESET
MS Code : SERRA_840 | CP | DISK

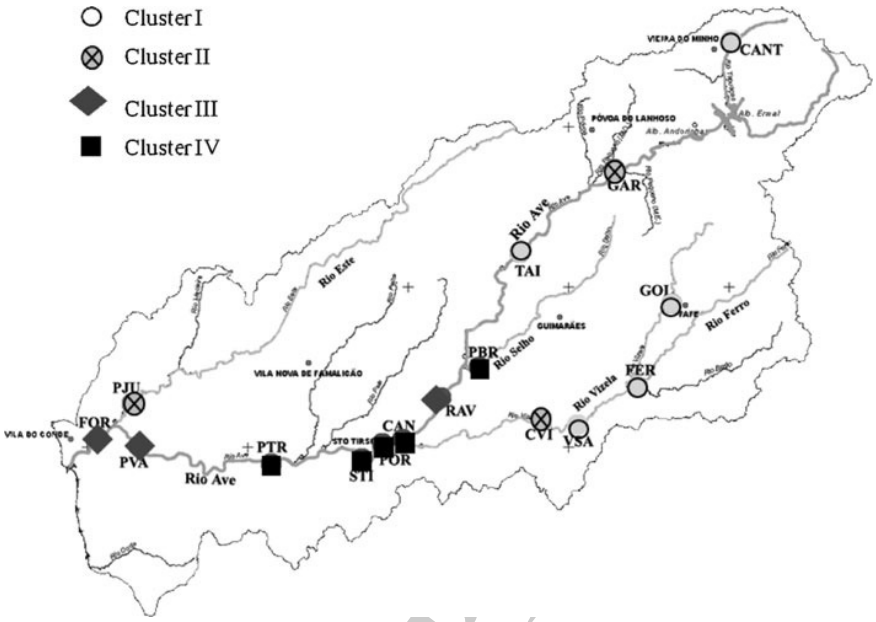

Fig. 4 Representation of clusters in the River Ave hydrological basin

make monthly predictions of the DO concentration, the seasonality component should not be omitted in models formulations.

4.1 Linear models

This section presents the linear models used to describe the main characteristics of the DO series and to adjust different models to each homogeneous group of monitoring sites (cluster). Linear models are primary tools in the context of environmental problems. There are many contributions (Carl and Kühn 2008; Mouriño and Barão 2009; Paschalidou et al. 2009). Linear models are simple and have good statistical properties; they are very robust statistical methods and this feature makes them a very attractive framework to describe the quality variables under study. It is well known that the choice of the independent variables should rely on the principle of parsimony. Also, their selection is heavily context dependent, because both variables themselves and the respective estimates for the coefficients should have a clear interpretation within the framework in which the case study is included.

Environmental data are naturally affected by the different seasons and by the environmental degradation that has been verified in a more aggravated way in recent years. Thus, any model to predict the behaviour of data must take these two factors into account. In this case study there is no measure of space continuity and, therefore, the observations at different locations in the cluster will be treated as independent observations, referenced in time. Hence, within each cluster we consider a variable observation by $Y_{j,t}^{(i)}$, where i represents the cluster, $i = 1, 2, 3, 4$, j represents the monitoring site running along all the sites in the cluster i, $j = 1, \ldots, k_i$ and $t = 1, \ldots, n_j^{(i)}$ stands for the month. With this notation, the model in each cluster i includes two additive components corresponding to different types of effects, that is,

$$Y_{j,t}^{(i)} = T_t^{(i)} + S_t^{(i)} + e_{j,t}^{(i)}. \quad (6)$$

Let us now analyse in more detail how to describe each component with the help of a linear model. The trend is generally described by a simple linear function of time, $T_t^{(i)} = \alpha^{(i)} + \beta^{(i)} t$. The seasonal component $S_t^{(i)}$ is a periodic function taking 12 different values, say, $\lambda_s^{(i)}$, $s = 1,\ldots, 12$ associated with each month of the year and expressing the positive or negative deviation from the trend due to the effect of that month. This type of effect is usually described with the help of 12 dummy variables indicating if each time instant t corresponds to month i. However, when the model has a constant term, in order for these parameters to be estimable they have to add up to 0, that is, $\sum_{s=1}^{12} \lambda_s = 0$ and $\lambda_{12}^{(i)} = -\sum_{s=1}^{11} \lambda_s^{(i)}$. Thus, as one of the seasonal coefficients has to be written as a function of the others, the seasonal component is represented by a linear combination of 11 explanatory variables

$$S_t^{(i)} = \sum_{s=1}^{11} \lambda_s^{(i)} S_{s,t},$$

$$\text{where } S_{s,t} = \begin{cases} 1, & \text{if date } t \text{ corresponds to month } s \\ -1, & \text{if date } t \text{ corresponds to month } 12 \\ 0, & \text{otherwise} \end{cases}.$$

Clearly, the choice of the twelfth month, December, to be written as a linear combination of the others is arbitrary

Journal : Large 477 | Dispatch : 20-7-2010 | Pages : 13
Article No. : 429 | LE | TYPESET
MS Code : SERRA_840 | CP | DISK

and any month can be used for that role. Finally, the model includes a stochastic component, $e_{j,t}^{(i)}$, which we suppose to be simply a white noise process, that is, a sequence of uncorrelated zero mean random variables, with constant variance $\sigma^{2(i)}$. A careful check of residuals shows that there were no significant violations of the normality and independence conditions. This procedure ensures the optimality properties of the OLS as well as the power of the t and F tests performed. After the model with all variables (full model) was adjusted, the authors used a backwards elimination procedure to select the significant variables. The regressor with largest p-value for its t-statistic was removed at each step, until all the regressors were significant at the level 0.05. The final reduced model was also tested against the full model with the help of a F-test on the set of all removed independent variables.

The DO modelling procedure starts by fitting the linear model (6) to the DO data observed in each cluster. The regression parameters obtained for each cluster are presented in Table 3. As expected, the four clusters show a seasonal pattern with lower values of DO concentration in the warmer months as compared to autumn and winter months. This could be expected because the inverse relationship between temperature and DO is a natural process—warmer water becomes more easily saturated with oxygen and it can hold less DO. Cluster I presents a weak positive significant trend associated to the sites in rural areas near the source of the river with good water quality. Cluster II and Cluster III present a weak decreasing trend associated to polluted areas that are densely populated, with high industrial productivity and where the Ave also receives similarly polluted waters from its adjacent streams. Cluster III presents the highest coefficient of determination (69%). Cluster IV, the most polluted cluster, has a stable behaviour with no significant trend, which may be justified if we take into consideration the highest variability when compared with other clusters. In this case, the coefficient of determination was 57%.

Table 3 Results for linear models adjustment to the four clusters

	Cluster I	Cluster II	Cluster III	Cluster IV
Intercept	9.197	9.867	8.634	7.711
Trend	0.004	−0.009	−0.003	–
January	0.960	1.382	2.045	2.540
February	1.117	1.410	2.122	2.457
March	0.602	0.577	0.983	1.323
April	–	0.784	0.708	1.027
May	–	–	–	–
June	−0.440	−1.093	−1.025	−1.352
July	−1.160	−1.557	−3.143	−3.391
August	−1.426	−2.037	−1.755	−1.501
September	−0.980	−0.912	−2.441	−3.240
October	−0.520	−0.546	−0.780	−1.143
November	0.720	0.711	1.183	1.118
December	1.125	1.281	2.103	2.161
$\hat{\sigma}^{2(i)}$	0.854	1.019	1.198	1.766
R^2	0.50	0.58	0.69	0.57

4.2 State space models

In the previous section it was verified that the seasonality is an important structural component to predict the monthly DO concentration. Thus, a regression model with varying coefficients (Pagan 1980; Leybourne 2006) represented in a state space framework could improve the predictions accuracy. However, in this case it implies establishing multivariate models that involve a large number of parameters, in addition to a complex structure that difficults its interpretation.

We propose an alternative model that assumes prior knowledge of DO seasonal coefficients s_t or its estimates (in this case, its monthly means computed in the past). This assumption implies some knowledge of the seasonal component behaviour. However, the modelling of this type of data is usually performed under previous studies or considering a significant data set. Another solution to overcome the monthly means of DO could be the use of seasonal coefficients estimates obtained via the linear process. Nevertheless, we selected the simplest option to compute the monthly means of observations, to compare linear regression with state space approaches in order to guarantee the independency of two estimation processes.

The model assumes that for month t the measurement $Y_{j,t}^{(i)}$ of the DO in monitoring site j of the cluster i is the calibrated seasonal coefficient added to a zero mean error, i.e., $Y_{j,t}^{(i)} = s_t^{(i)}\beta_t^{(i)} + e_{j,t}^{(i)}$, where $\beta_t^{(i)}$ is a calibration factor of cluster i at time t. Considering that cluster i is composed by k_i water monitoring sites, the DO can be modelled at cluster i by

$$\mathbf{Y}_t^{(i)} = \mathbf{s}_t^{(i)}\beta_t^{(i)} + \mathbf{e}_t^{(i)}. \quad (7)$$

$$\beta_t^{(i)} - 1 = \phi^{(i)}\left(\beta_{t-1}^{(i)} - 1\right) + \varepsilon_t^{(i)}. \quad (8)$$

where $\mathbf{Y}_t^{(i)} = \left[Y_{1,t}^{(i)}, Y_{2,t}^{(i)}, \ldots, Y_{k_i,t}^{(i)}\right]'$ and $\mathbf{s}_t^{(i)} = s_t^{(i)} \cdot \mathbf{1}_{k_i}$. The error vector $\mathbf{e}_t^{(i)} = \left[e_{1,t}^{(i)}, e_{2,t}^{(i)}, \ldots, e_{k_i,t}^{(i)}\right]'$ and the error $\varepsilon_t^{(i)}$ are zero means uncorrelated errors, $E\left[e_{j,t}^{(i)} e_s^{(i)}\right] = 0$ for all t, s and j, with matrix of covariance $E\left[\mathbf{e}_t^{(i)}\mathbf{e}_t^{'(i)}\right] = \sigma_{e,i}^2 \cdot \mathbf{I}_{k_i}$ and variance $E\left[\varepsilon_{t,i}^2\right] = \sigma_{\varepsilon,i}^2$, respectively. The calibration factor $\beta_t^{(i)}$ is assumed to be a stationary autoregressive process of order 1, $\left|\phi^{(i)}\right| < 1$, with unitary mean.

Journal : Large 477 Dispatch : 20-7-2010 Pages : 13
Article No. : 429 LE TYPESET
MS Code : SERRA_840 CP DISK

Table 4 Seasonal coefficients (monthly means) of the DO concentration of the four clusters

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Cluster I	10.66	10.87	10.25	9.89	9.82	9.14	8.48	8.36	8.64	9.44	10.34	10.77
Cluster II	10.97	10.53	9.96	9.97	9.12	8.47	7.52	7.25	8.06	8.43	9.76	10.63
Cluster III	10.22	10.24	9.16	8.92	8.52	7.21	5.16	6.47	5.80	7.46	9.33	10.09
Cluster IV	10.10	10.03	8.98	8.70	8.21	6.56	4.11	6.28	4.36	6.62	8.80	9.88

Table 4 summarizes seasonal coefficients estimates for the four clusters computed by the mean of the observed values of the DO concerning monitoring sites that belong to each cluster and according each month. Table 5 presents gaussian maximum likelihood estimates of parameters (referred in Sect. 3.2) that concern each cluster's multivariate model (7)–(8). As mentioned before, the state process $\beta_t^{(i)}$ can be interpreted as a dynamic calibration factor of the seasonal component; so, it is reasonable that state noise has a small variance, as can be seen in Table 5. The largest values of both state and observation equations error variances are obtained in Clusters III and IV. As mentioned before concerning the clustering analysis, these clusters include the water monitoring sites which have the lowest annual DO values. Therefore, Clusters III and IV include sites with worse indicators in what concerns DO concentration, but at the same time they also present the largest variability. Nevertheless, Cluster IV presents worse water quality indicators that clearly distinguish it, even from Cluster III.

Taking advantage of the interpretability of the state process, Fig. 5 shows filtered state values estimates $\hat{\beta}_{t|t}$ in the modelling period. As the graphic shows, the model captures changes in a dynamic way that overlaps the default behaviour evidenced by the seasonality. The state space model (7)–(8) provides a useful tool to evaluate changes in real time on DO concentration in each month. Indeed, calibration factor estimates greater than one indicate an improvement in water quality, while estimates lower than one can indicate water quality deterioration. This possibility of model formulation (7)–(8) benefits the water monitoring process in a way that linear models are not able to provide, since linear models incorporate global trends in pre-established time periods.

Table 5 Estimated values obtained for the multivariate models parameters of the four clusters via gaussian maximum likelihood estimation

Cluster	$\hat{\phi}$	$\hat{\sigma}_\varepsilon^2$	$\hat{\sigma}_e^2$
I	0.5213	0.0016	0.3140
II	0.5246	0.0035	0.5549
III	0.7527	0.0036	0.7138
IV	0.2242	0.0158	1.6322

4.3 Comparative analysis

A first analysis of adjustment models is done comparing data and predictions. Globally, models fit the data satisfactorily. For instance, Fig. 6 shows the observed values and one-step predictions of the DO concentration concerning Cluster I from November 1992 to September 2004, with 95% point-wise prediction intervals. Graphic representation shows that the linear model produces point-wise prediction intervals with large amplitude than state space models, as is the case in other clusters. As the state space model incorporates the last observation at each time, it enables updating and accurate predictions.

Accurate forecasts are important information on water monitoring processes; therefore, they are a good criterion to assess the model performance. Thus, we compare the linear and state space models performance concerning the one-step forecast's mean square error, relatively to the remaining unused data in the modelling process. So, we computed forecasts for the 228 observations by using two approaches. However, approximately half of this data is related to Cluster I, because some water monitoring sites belonging to other clusters were inactive due to public policy restructuring.

In order to compute forecasts based on linear models, we applied the estimated regression equation obtained from Table 3, while state space models forecasts are evaluated through Kalman filter predictors. The Kalman filter algorithm produces the best linear prediction $\hat{\beta}_{t|t-1}^{(i)}$ for the calibration factor $\beta_t^{(i)}$ for month t and cluster i and, consequently, the best linear prediction of DO is $\hat{Y}_t^{(i)} = s_t^{(i)}\hat{\beta}_{t|t}^{(i)}$, which is clearly equal for all sites from a same cluster at each month.

In Fig. 7 we present forecasts with 95% point-wise prediction intervals of DO concentration from October 2004 to October 2006. We do not have records from October 2005 in any of the water monitoring sites and we were not able to determine the cause of this occurrence. This representation indicates that state space approach allows obtaining prediction intervals with lower range than linear models. The state space model tends to fit data better, whereas the linear model seems to overestimate the values. In this concrete cluster, the linear model considers the trend as a significant component with a small positive slope in the modelling period. Nevertheless, its rigid

Journal : Large 477 Dispatch : 20-7-2010 Pages : 13
Article No. : 429 LE TYPESET
MS Code : SERRA_840 CP DISK

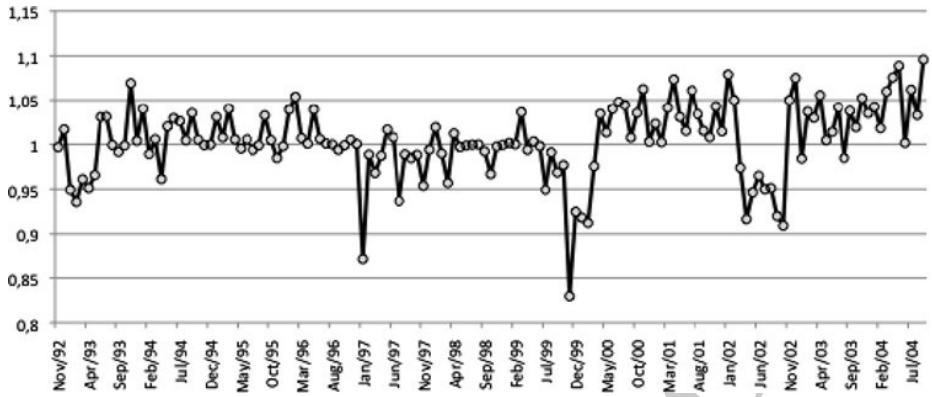

Fig. 5 Filtered state values estimates $\hat{\beta}_{t|t}$ for cluster I

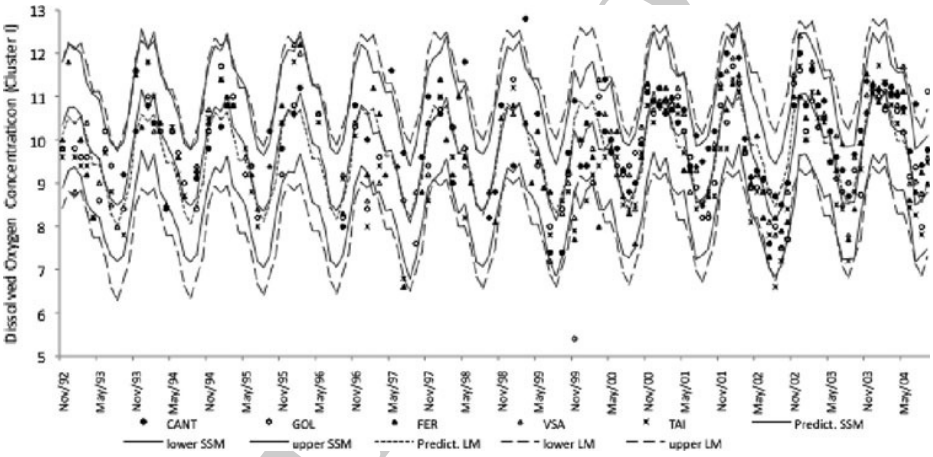

Fig. 6 Observed values of the DO concentration for the Cluster I and one-step predictions by linear model and state space model approaches from November 1992 to September 2004, with 95% point-wise prediction intervals

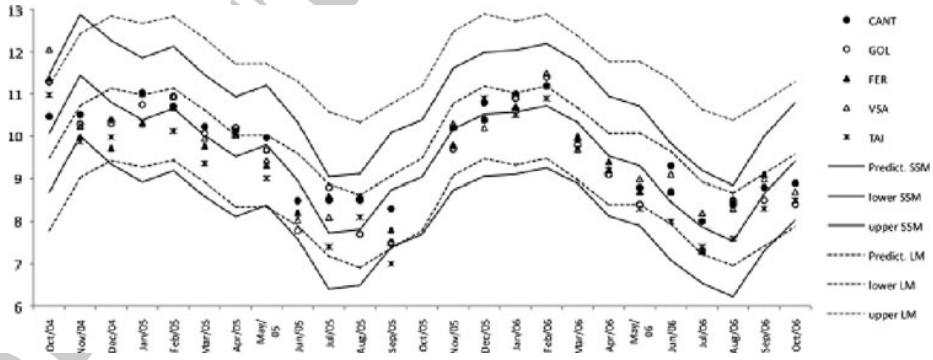

Fig. 7 One-step prediction with 95% point-wise prediction intervals of the DO concentration from October 2004 to October 2006 (Cluster I) by linear and state space models

structure may not have detected temporary changes or new conditions such as new industries in that area, for instance.

In order to compare the two approaches in a quantitatively point of view, in the sense of predictions accuracy, we computed the root of the mean square errors (RMSE) of predictions, i.e.,

$$\mathrm{RMSE} = \sqrt{\frac{1}{228}\sum_{i=1}^{4}\sum_{t=1}^{n_i}\left(Y_t^{(i)} - \hat{Y}_t^{(i)}\right)^2}.$$

Taking into account all 228 predictions computed in the period of the assessment procedure, we obtained RMSE = 0.961 in the linear models and RMSE = 0.846 in the state space approach. Globally, it is the state space modelling approach that provides the least RMSE as far as concerns all the predictions we made. So, the dynamic structure of the state space models improved the predictions accuracy in the sense of the mean square error.

5 Conclusions

This work shows that state space approach combined with clustering techniques allows identifying homogeneous

Journal : Large 477 | Dispatch : 20-7-2010 | Pages : 13
Article No. : 429 | LE | TYPESET
MS Code : SERRA_840 | CP | DISK

groups of water monitoring sites, based on similarities in the temporal dynamics of monthly records of DO concentration. With an appropriate disparity measure based on Kullback information, we identified four clusters of water monitoring sites in the River Ave basin that allow us to construct only four models in order to forecast this variable in the future by reducing the initial 16 water monitoring sites. Besides, it is possible to arrange these groups by order of water quality degree, from the less polluted (Cluster I) to the most polluted (Cluster IV). So, the less polluted monitoring sites are near the sources of River Ave or its adjacent streams (except for river Selho), and the most polluted water monitoring sites are located in a highly industrial area exposed to discharges of industrial effluents.

After a clustering procedure, the comparison between the forecast's mean square error of the linear model and of the state space model shows that the latter evidences an improved accuracy within the proposed assessment period. We adopted a state space model to predict the monthly DO concentration, which calibrates the seasonal coefficients through an autoregressive calibration factor. This approach is an alternative to the standard linear model with tendency and seasonality components, because this model incorporates a dynamic structure that allows an easy data fitting. Furthermore, calibration factors have a useful interpretation as an approximate ratio between observed measure of the DO concentration and the respective seasonal coefficient. This approach provides a real time procedure to monitoring DO concentration in which calibration factors greater or lower than one indicate that water quality improved or deteriorated in comparison to the expected value based on past behaviour. From the forecast point of view, the state space model reduced the forecast's mean square error from 0.961 (obtained with linear model) to 0.846, which is a significant improvement.

We hope that this work could be a tool for decision support, because monitoring procedures and good models of water quality variables are indispensable, mainly in a highly industrial region as is the Ave Valley in the northwest of Portugal.

Acknowledgments The authors would like to thank the anonymous referees for many helpful critics and suggestions that contributed to improve this paper. The authors would like to thank to Eng. Pimenta Machado from the Portuguese Regional Directory for the Northern Environment and Natural Resources and to Eng. Cláudia Brandão from the Portuguese Institute of Water, for sharing their skills and experiences and for supplying the monitored data. A. Manuela Gonçalves acknowledges the financial support provided by the Research Centre of Mathematics of the University of Minho through the FCT Pluriannual Funding Program.

References

Alpuim T, Barbosa S (1999) The Kalman filter in the estimation of area precipitation. Environmetrics 10:377–394

Bengtsson T, Cavanaugh JE (2008) State-space discrimination and clustering of atmospheric time series data based on Kullback information measures. Environmetrics 19:103–121

Boi P (2004) A statistical method for forecasting extreme daily temperatures using ECMWF 2-m temperatures and ground station measurements. Meteorol Appl 11:245–251

Brown P, Diggle P, Lord M, Young P (2001) Space-time calibration of radar rainfall data. Appl Stat 50(2):221–241

Carl G, Kühn I (2008) Analysing spatial ecological data using linear regression and wavelet analysis. Stoch Environ Res Risk Assess 22(3):315–324

Costa M, Alpuim T (2010) Parameter estimation of state space models for univariate observations. J Stat Plan Inference 140(7):1889–1902

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B 39:1–38

Everitt BS, Landau S, Leese M (2001) Cluster analysis, 4th edn. Arnold, London

Fovell R, Fovell M (1993) Climate zones of the conterminous United States defined using cluster analysis. J Clim 6:2103–2135

Galanis G, Anadranistakis M (2002) A one-dimensional Kalman filter for the correction of near surface temperature forecast. Meteorol Appl 9:437–441

Gong X, Richman M (1995) On the application of cluster analysis to growing season precipitation data in North America east of the Rockies. J Clim 8:897–931

Harvey AC (1996) Forecasting structural time series models and the Kalman filter. Cambridge University Press, Cambridge

Kullback S (1968) Information theory and statistics. Dover, New York

Leybourne SJ (2006) Estimation and testing of time-varying coefficient regression models in the presence of linear restrictions. J Forecast 12(1):49–62

Libonati R, Trigo I, DaCamara C (2008) Correction of 2 m-temperature forecasts using Kalman filtering technique. Atmos Res 87:183–197

Mouriño H, Barão MI (2009) A comparison between the linear regression model with autocorrelated errors and the partial adjustment model. Stoch Environ Res Risk Assess 24(4): 499–511

Oliveira RES, Lima MMCL, Vieira JMP (2005) An indicator system for surface water quality in river basins. In: Inter-Celti colloquium on hydrology and management of water resources 4, Guimarães

Pagan A (1980) Some identification and estimation results for regression models with stochastically varying coefficients. J Econom 13:341–363

Paschalidou AK, Kassomenos PA, Bartzokas A (2009) A comparative study on various statistical techniques predicting ozone concentrations: implications to environmental management. Environ Monit Assess 148:277–289

PGIRH/N (1988) Metodologias para a Avaliação de Políticas de Recursos Hídricos - Plano de Gestão da Bacia Hidrográfica do Rio Ave (in Portuguese). Ministério das Obras Públicas, Transportes e Comunicações, Laboratório Nacional de Engenharia Civil, Ministério do Plano e Administração do Território, Comissão de Coordenação da Região Norte 8:66, Lisboa

PGIRH/N and NATO PO-RIVERS (1994) Caracterização e Directrizes de Planeamento dos Recursos Hídricos do Norte – A Bacia Hidrográfica do Rio Ave (in Portuguese). Ministério do

Journal : Large 477 | Dispatch : 20-7-2010 | Pages : 13
Article No. : 429 | LE | TYPESET
MS Code : SERRA_840 | CP | DISK

Ambiente e dos Recursos Naturais, Direcção Regional do Ambiente e Recursos Naturais, Instituto da Água. Porto 1–5, 1–13

Shrestha S, Kazama F (2007) Assessment of surface water quality using multivariate statistical techniques: a case study of the Fuji river basin, Japan. Environ Model Softw 22:464–475

Shumway R, Stoffer D (1982) An approach to time series smoothing and forecasting using the EM algorithm. J Time Ser Anal 3:253–264

Shumway R, Stoffer D (2006) Time series analysis and its applications, 2nd edn. Springer-Verlag, Berlin

Stone RC (1989) Weather types at Brisbane, Queensland: an example of the use of principal components and cluster analysis. Int J Climatol 9:3–32

Vieira JMP (2003) Water management in national water plan challenges (in Portuguese). Revista Engenharia Civil 16:5–12

Zhu R, El-Shaarawi AH (2009) Model clustering and its application to water quality monitoring. Environmetrics 20:190–205

Journal : Large 477 | Dispatch : 20-7-2010 | Pages : 13
Article No. : 429 | ☐ LE | ☐ TYPESET
MS Code : SERRA_840 | ☑ CP | ☑ DISK